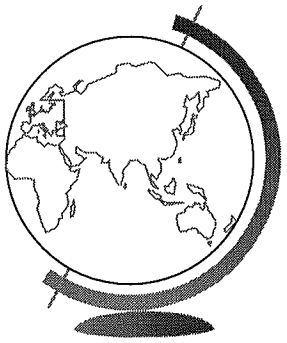
File
Presentations
FDIS talk.
Key West.



Archiving the Internet: Towards a Core Internet Service

Brewster Kahle
President, Internet Archive
brewster@archive.org
April 26, 1996

What is it?
Who will care?
Is it Possible?
Why do we want to do it?

Other Repositories

- Library of Alexandria: 800GB (400k scrolls @2MB)
- Library of Congress: 20TB (20M books, ascii)
- Dialog Information Service: 3-5TB
- Video Store: 8TB (5k videos, 1GB/hr)
- Public Branch Library: 3TB (300k scanned books)
- Radio Station: 1TB (15k hrs of music)
- . . . Internet Archive: 1-10TB

2

Our Mission is to . . .

- Gather, Archive, and Serve all public Internet information (WWW, Netnews, Gopher, and Usage Logs)

3

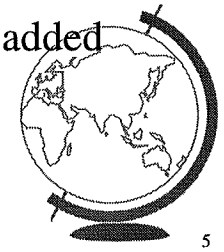
Offering for the first time . . .

- Reliability (Backing store for Net resources)
- Accountability (Official copy of record)
- Durability (Library for Internet research community)
  - Demographics, clustering, indexing

4

Who Will Care?

- Users: reliable access to the Net resources (built into browsers and proxies)
- Scholars/Historians: understanding the new medium
- Marketers: demographic treasure-drove
- Entrepreneurs: basis for new value-added services



5

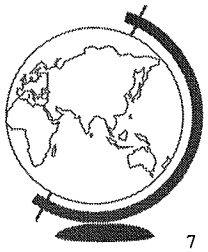
Is it Doable?

- Legal/Social Issues:
  - Privacy
  - Copyright/licensing
  - Export/Pornography
- Technical:
  - Gathering
  - Storage
  - Access

6

Gathering

- Methods: Crawling, Tape donations, Satellite receiver
- Technology: Tuned machines, mostly custom software
- Speed: T3(45Mb/s) = 500 GB/day, 66¢/GB



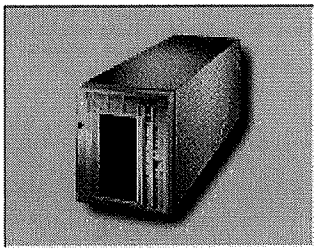
7

Storage

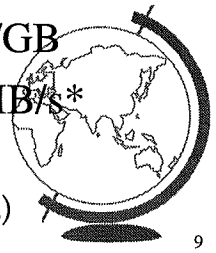
- Mitra's Law of Archiving: For every dollar they spend, we can only spend a nickel
- Disks: $200/GB, RAID: $500/GB
- Luckily, tape costs recently plummeted

8

Storage: DLT 7700



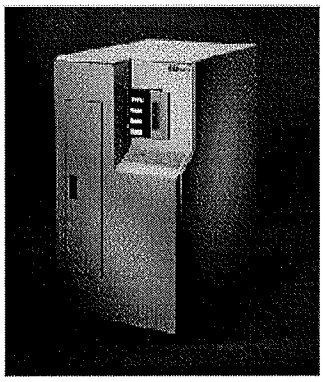
- # tapes: 7
- # drives: 1
- Storage: 490GB*
- Cost: $11k
- Cost/GB: $23/GB
- Speed: 10MB/s*



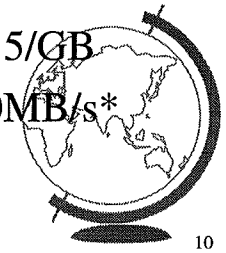
* Compressed (for native, divide by 2)

9

Storage: ATL Odetics 452



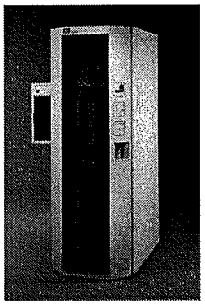
- # tapes: 52
- # drives: 4
- Storage: 3.6TB*
- Cost: $54k
- Cost/GB: $15/GB
- Speed: 40MB/s*



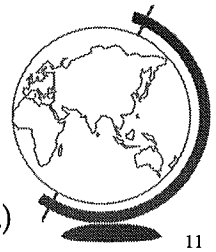
* Compressed (for native, divide by 2)

10

Storage: ATL Odetics 2640



- # tapes: 264
- # drives: 3
- Storage: 18TB*
- Cost: $100k
- Cost/GB: $6/GB*
- Speed: 30MB/s*



* Compressed (for native, divide by 2)

11

Therefore, We have the Technology

- Gathering 10TB takes 20 days, $10k raw bandwidth, custom software (+ CPUs)
- Storing 10TB takes $100k for robot, ingenuity (+ fast I/O)
- Public Access takes calculations and fresh ideas

12

Where does the Technology Lead?

- Intranet applications
- Video Storage/Servers
- Data mining
- International Internet Centers
- Towards an "Internet Operating System" of backup, cache consistency, accounting, directory, file storage . . .

13

Impact of the Archive

- Transition the Net from Ephemera to an Enduring Medium
- Inject extra computing services for navigation, reliability, coordination
- Build lasting position IN the Net (not ON the Net)

14

Building a Library that can Think.

What does it take?

- ✦ Bandwidth
- ✦ Computes
- ✦ Smarts
- ✦ Gumption

15

617-478-2332

570-465-2750